Partial-Least-Square PLS und Kennzahl VIP

Voraussetzung und verwandte Themen

Für diese Beschreibungen sind Grundlagen der Statistik vorteilhaft. Weiterführende und verwandte Themen sind:

www.versuchsmethoden.de/Multiple_Regression.pdf

www.versuchsmethoden.de/Hauptkomponentenanalyse.pdf

Einführung

PLS wurde 1960 von dem schwedischen Ökonometriker Herman Wold entwickelt. PLS steht für: „Partial Least Squares Modeling in Latent Variables“. Der Zweck ist vor allem die Auswertung von korrelierenden Daten oder von Mischungsplänen, bei denen die Multiple Regression (MR) nicht anwendbar ist. Ein wesentlicher Vorteil von PLS ist auch, dass viele Variablen verarbeitet werden können. Es ist sogar möglich mit weniger „Messdaten“ für die Auswertung auszukommen, als Variablen vorhanden sind.

Ziel und Nutzen

PLS hat sich in den Bereichen Pharma, Chemie und Spektroskopie als Standard durchgesetzt. Häufig wird PLS als Universalmethode für alle Auswertungen gesehen. Für Auswertung von Daten die nicht zu stark oder gar nicht korrelieren (z.B. aus der Versuchsplanung) ist aber nach wie vor die Multiple Regression vorzuziehen, da hier die Effekte und Modelle besser zu interpretieren sind. Bei rein orthogonalen Daten sind allerdings die Koeffizienten der Regressionsmodelle auch gleich.

Grundlagen

PLS ist mit der Hauptkomponentenanalyse PCA (*Principal Component Analysis*) sehr verwandt. Im Gegensatz zu PCA gilt hier der Zusammenhang mit der Gewichtsmatrix W anstelle der Loadings:

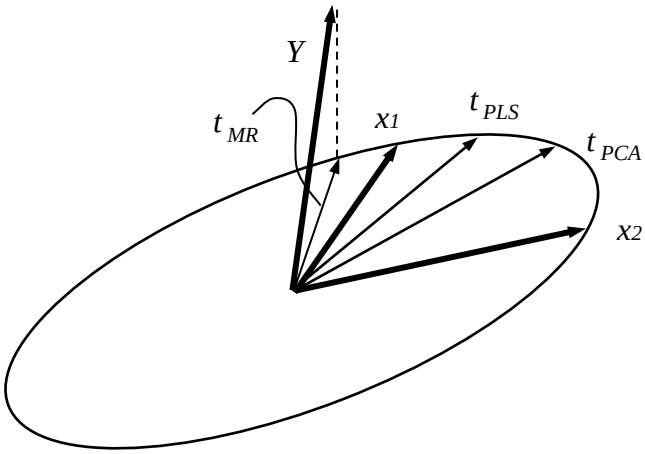
$$X = T W^T$$

In W ist die Zielgröße y enthalten, die es in der PCA nicht gibt. Auch hier gilt

$$w_1^2 + w_2^2 + ...w_k^2 = 1$$

Das Regressionsmodell ist:

$$\hat{y} = T c^T$$



Wobei c die Regressionskoeffizienten darstellt. Das dargestellte Bild zeigt zwei Variablen x_1 und x_2. Die Hauptkomponentenanalyse PCA mit t_{PCA} liegt in der „Beule“ der Ellipse, die sich in der Ebene von x ergibt. Je stärker x_1 und x_2 korrelieren, desto länger wird t_{PCA}. Liegt keine Korrelation vor, ist die Vektorrichtung von t_{PCA} nicht mehr definiert, denn die Ellipse wird dann zu einem Kreis und hat keine Vorzugsrichtung mehr. Die Komponente t_{PLS} dagegen ist über die Betrachtung der Kovarianz dann immer noch bestimmbar. Das ist ein entscheidender Vorteil von PLS

gegenüber PCA. Die Ergebnisse, d.h. die ermittelten Koeffizienten der Variablen sind dann identisch mit der MR-Methode (für orthogonale Daten). Wärend die MR-Methode bei hochkorrelierenden Daten nicht mehr eindeutige Ergebnisse liefert oder ganz aussteigt, kann die PLS-Methode weiterhin angewendet werden. Selbst wenn zwei Variablen zu 100% korrelieren ist das noch möglich. Natürlich ist die Zuordnung der Effekte dann nicht mehr eindeutig, PLS vergibt in diesem Fall den beiden Variablen jeweils den halben Effekt zu gleichen Anteilen.
Der Nachteil des PLS-Verfahrens ist, dass die Prognosen und R^2 schlechter sind als bei MR. Auch sind die Koeffizienten teilweise wesentlich kleiner, was dazu führt die Effekte zu gering zu schätzen.
Der komplette Algorithmus (NIPALS – *Nonlinear Iterative Partial Least Square*) stellt sich wie folgt dar:

$$w' = \frac{X^T y}{y^T y}$$ *Wichtungen absolut für standardisierte Matrix X*

$$w = w' / \sum w'^2$$ *Wichtungen normiert*

$$t = Xw$$ *Score Vektor*

$$= \frac{\sum_{j=1}^{z} \mathrm{cov}(y, x_j) x_j}{\sum_{j=1}^{z} \mathrm{cov}(y, x_j)^2}$$ *mit z = Anzahl Variable*

$$c = \frac{y^T t}{t^T t}$$ *Regressionskoeffizienten zwischen y und Komponente*

$$p = \frac{X^T t}{t^T t}$$ *Ladungs-Vektor*

$$E = X - tp^T$$ *Residuen-Matrix der Variablen*

$$f = y - tc^T$$ *Residuen-Vektor der Zielgröße*

Die nächsten Komponenten werden bestimmt, indem man $X{=}E$ und $y{=}f$ setzt und von vorne berechnet. Die Hinzunahme weiterer Komponenten erhöht in der Regel R². Ist das nicht der Fall, so werden keine weiteren Komponenten benötigt. Durch weitere Komponenten kann es passieren, dass einzelne Koeffizienten gravierend ihre Größe ändern. Das Modell mit der größeren Anzahl Komponenten ist dann maßgebend.

Bezogen auf den Regressionsansatz zwischen y und den ursprünglichen Variablen errechnen sich die Koeffizienten b über:

$$b = W(P^T W)^{-1} c^T$$

Zusammenfassende Eigenschaften:

- R^2_{PLS} ist kleiner R^2_{MR}
- Koeffizienten PLS sind kleiner als bei MR ⇨ Fehler wirken sich dadurch auch geringer aus.
- PLS maximiert die Kovarianz zwischen Hauptkomponenten und Y,

MR maximiert dagegen die Korrelation zwischen X und Y

- PLS kann mit hohen Korrelationen zwischen den Variablen X umgehen.

Schätzung der Streuung

Die Streuung der Koeffizienten b kann hier nicht grundsätzlich, wie bei der MR-Methode aus der Spur von $(X^T X)^{-1}$ berechnet werden. Ist die Korrelation zwischen den Variablen groß, so kann eine Streuung nur über eine so genannte Kreuzvalidierung erfolgen. Die häufigst verwendete Variante läßt sich durch folgendes Verfahren beschreiben: Jede Zeile aus der Matrix X wird einmalig herausgelassen (one leave out). Mit der reduzierten Matrix wird das Modell mit seinen Koeffizienten mit PLS berechnet. Es ergeben sich somit n unterschiedliche b. Hieraus kann man eine Standardabweichung für b bestimmen und somit die Streuung schätzen. Dabei wird klar, dass die ermittelte Standardabweichung stark von der Anzahl der Versuche abhängig ist, was bei geringem Umfang kritisch sein kann. Auch liefern Varianten der Kreuzvalidierung, z.B. bei Herauslassen von gleichzeitig mehreren Zeilen mit unterschiedlichen Umfängen, unterschiedliche Werte. Die Berechnung und Anwendung der p_{value}, wie beim MR-Verfahren ist hier also nicht zu empfehlen. Deshalb findet man für die PLS-Methode stattdessen häufig die VIP-Kennzahl zur Variablenselektion.

Variablenselektion mit VIP

Für das PLS-Verfahren eignet sich zur Variablenselektion die VIP-Kennzahl. VIP steht für ***V**ariable **I**mportance in the **P**rojection*, also wie wichtig der Einfluss der Variable in der Projektion auf die Scores t ist. Diese Kennzahl wurde erstmals 1993 von Wold veröffentlicht. VIP berechnet sich für die jeweilige Variable x_j über:

$$VIP_j = \sqrt{z \sum_{k=1}^{h} \left(\frac{y^T t_k}{t_k^T t_k} w_{jk}^2 \right) \Big/ \sum_{k=1}^{h} \left(\frac{y^T t_k}{t_k^T t_k} \right)}$$

mit h = Anzahl Komponenten,
z = Anzahl Ausgangsvariable (bzw. Terme)

Der y-Vektor muss hier standardisiert sein. In der Literatur wird für die VIP-Zahl ein Grenzwert zwischen 0,8 …1 genannt. Zu kleine Werte bedeuten, dass die Variablen aus dem Modell weggelassen werden sollten. Die Erfahrungen zeigen jedoch, dass auch VIP-Werte unter 1,0 nicht ungewöhnlich sind für Variable, die von Ihrem Einfluss auf das Modell trotzdem wichtig sind. Aus vielen Untersuchungen hat sich folgende Empfehlung ergeben, wenn ein Term im Modell bleiben sollte:

Variablen (Haupteffekte)	$VIP \geq 0{,}5$
Wechselwirkungen	$VIP \geq 1{,}0$

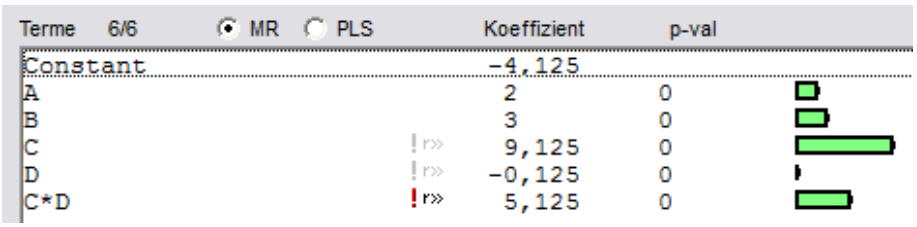
Bei der Frage, ob eine Variable aus dem Modell genommen werden sollte, ist auch die Größe des Einflusses (Koeffizient) zu berücksichtigen. Besonders von Bedeutung ist das Wissen um die tatsächlichen physikalischen Zusammenhänge als wichtigste Entscheidungsgröße.

A	B	C	D	Y
-1	-1	1	1	5
1	1	-1	-1	-3
-1	1	-1	-1	-7
1	1	-1	-1	-3
1	-1	-1	-1	-9
-1	-1	-1	-1	-13
1	1	1	1	15
-1	-1	1	1	5
-1	1	1	1	11
1	-1	1	-1	-1

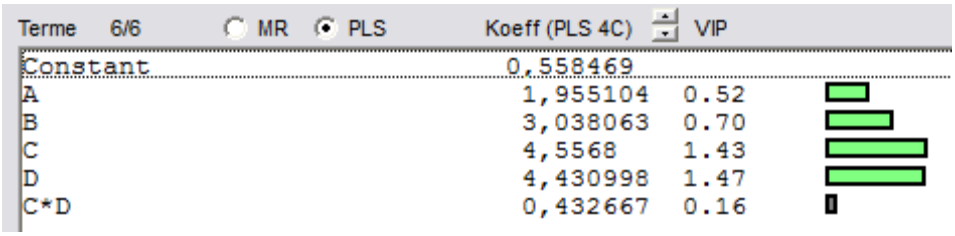
Partial-Least-Square PLS und Kennzahl VIP

Beispiel:

Im Datenbeispiel aus dem Kapitel ***Probleme mit zu stark korrelierenden Daten*** ergab sich nach Veränderung von Zelle D7 eine Korrelation zwischen Variable C und D von r = 0,816 (nur die letzte Zeile unterscheidet sich).
Die Methode MR zeigt fälschlicherweise eine hochsignifikante Wechselwirkung zwischen C und D an (p-val = 0), obwohl eine Wechselwirkung in der vorgegebenen Funktion nicht definiert war. Die wahren Koeffizienten C = 4 und D = 5 werden auch völlig falsch geschätzt.



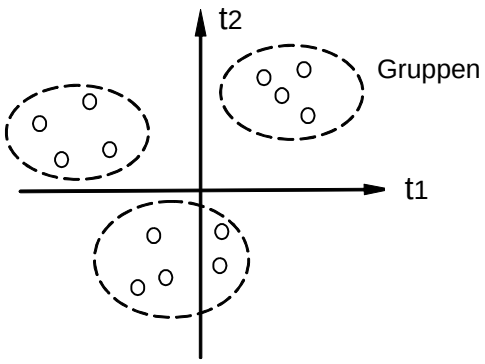
PLS ergibt zum Vergleich mit 4 Komponenten (4C) folgendes Bild (Hinweis: Die Daten wurden nicht standardisiert, also auf die Standardabweichung bezogen, sondern für den Vergleich auf -1 .. +1 normiert).

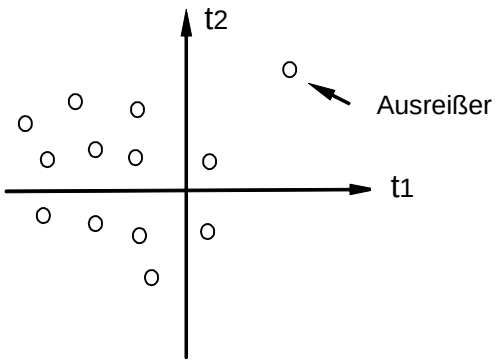


Die VIP-Zahl zeigt für die Wechselwirkung C*D korrekterweise an, dass sie nicht real ist (Empfehlung für Wechselwirkungen: VIP sollte ≥ 1 sein). Der wahre Wert von 1 für Constant wird in PLS ebenfalls besser geschätzt, als bei MR (-4,125).
Dies ist ein großer Vorteil von PLS. Die Schätzwerte für C und D werden wegen der Korrelation zwar als fast gleich ausgegeben, sind aber im Gegensatz zu MR relativ gut.

Score Plot

Der Score Plot stellt jeden Messpunkt über die wichtigsten Scores t_1 und t_2 dar. Dabei können evtl. Muster und Gemeinsamkeiten erkannt werden. Liegen Messpunkte eng zusammen, desto ähnlicher sind sie sich und umgekehrt. Dabei können auch markante Ausreißer erkannt werden.

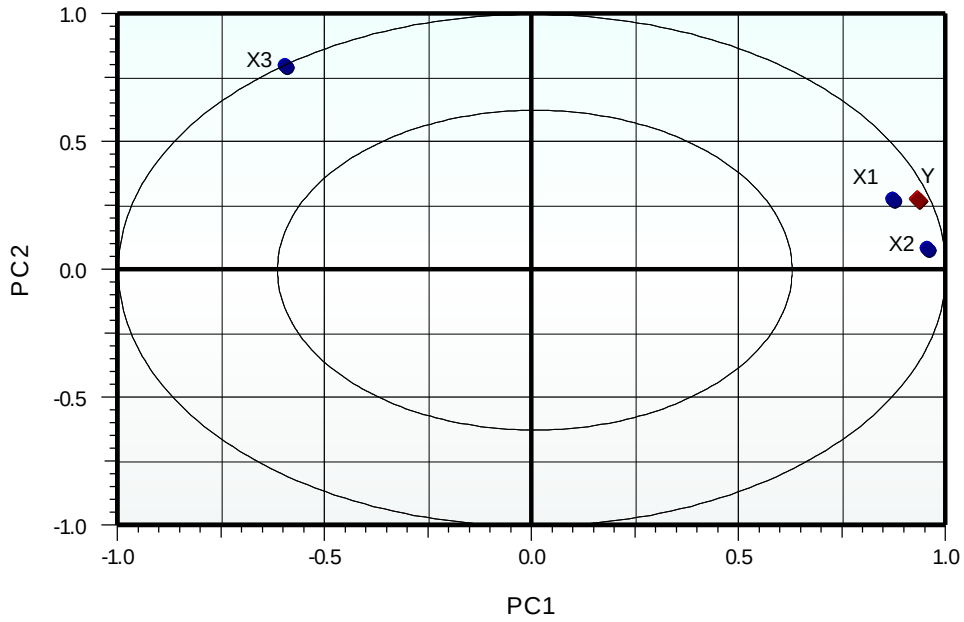




Partial-Least-Square PLS und Kennzahl VIP

Korrelations-Ladungen (Correlation Loading Plot)

Im so genannten Correlation Loading Plot werden die erklärten Varianzen der Variablen und der Zielgröße auf die Komponenten PC dargestellt. Die Achsen sind skaliert als Korrelationswerte wobei gilt: Erklärte Varianz = Korrelation². In diesem Diagramm werden die Einflüsse der Variablen aufgezeigt und man erkennt welche Komponenten besser die Variablen beschreiben.
Die Ellipsen stellen jeweils 100% (äußere) und 50% (innere) erklärte Varianz dar.



Je näher die Variablen an der 100% Ellipse liegen, desto wichtiger sind diese. In diesem Beispiel erklärt die Komponente PC1 die Variablen x1, x2, sowie die Zielgröße y fast alleine, während für x3 beide Komponenten notwendig sind.

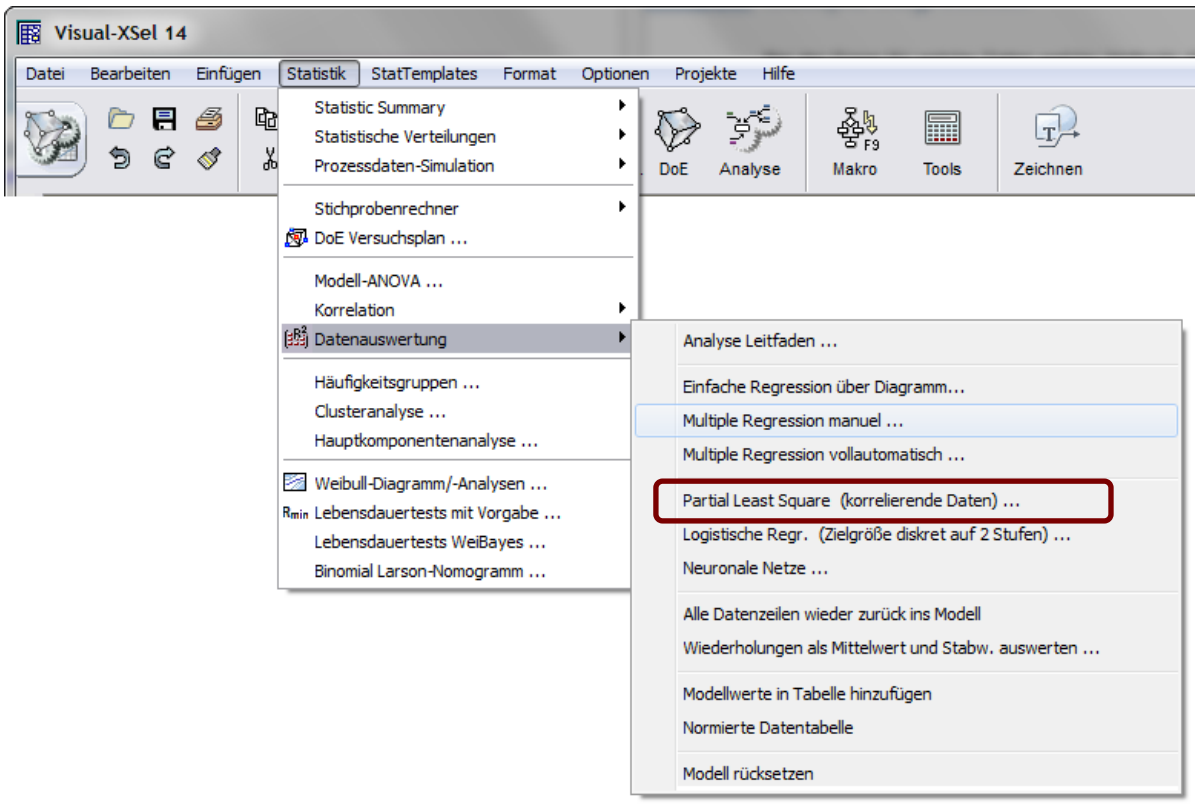
Anwendung in Visual-XSel 14.0

www.crgraph.de

Verwenden Sie für den Einstieg die Datenauswertung im Leitfaden,

oder die Ikone

oder den Menüpunkt Statistik….

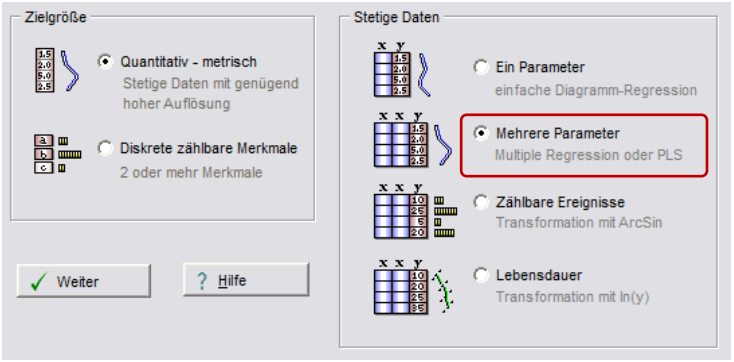


Die folgende Beschreibungen bezieht sich auf die Daten (Version 14.0) unter:

…\Beipieldaten\Beispiel_Verbrauch.vxt

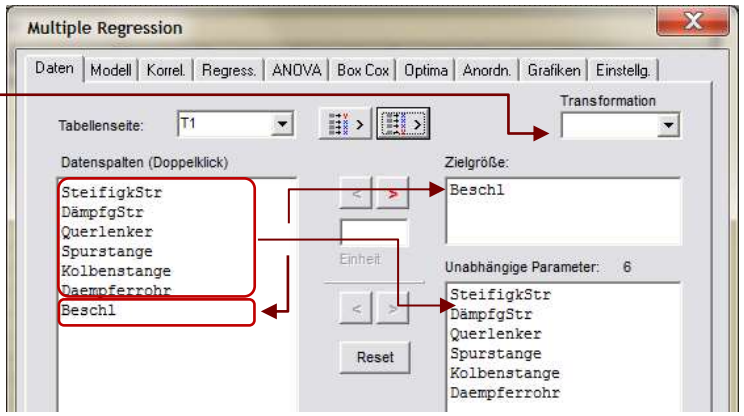
Partial-Least-Square PLS und Kennzahl VIP

Es erscheint zunächst eine Auswahl was für eine Art der Zielgröße vorliegt. Die Standardeinstellung ist stetige Messgrößen. Bei zählbaren Ereignissen wird math. So transformiert, dass das Ergebnis nur zwischen 0…1 liegen kann. Bei Lebensdauerwerten müssen diese logarithmiert werden und Partial-Least-Square kann mit korrelierenden Daten umgehen, die nicht aus einer DoE stammten. Die zuletzt genannte Methode kann man auch später auswählen, da die Korrelation erst noch zu prüfen ist.

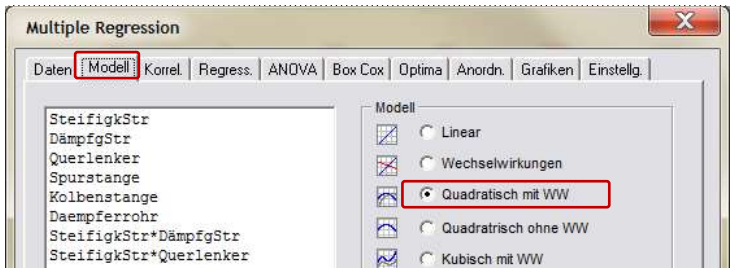


Ordnen Sie die in der Kopfzeile der Tabelle vorkommenden Titel jeweils als Zielgröße oder als Parameter zu.

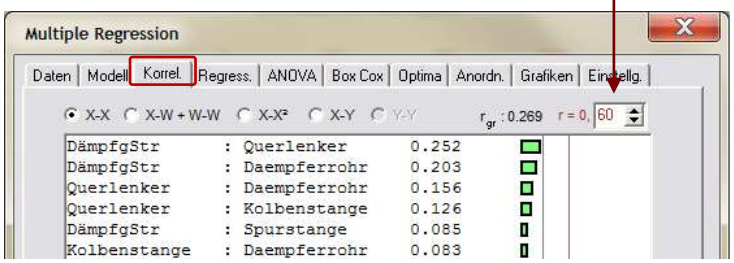
Sowohl Zielgröße, als auch Parameter können hier transformiert werden.
Wurde vorher die Zielgröße als Lebensdauer angegeben, steht hier ln(y).
Bei ungekannten math. Zusammenhängen hilft später als Entscheidung die Box-Cox Analyse.



Unter dem Reiter Modell kann ausgewählt werden, ob ein rein lineares Modell, mit Wechselwirkungen oder nichtlinearen Verläufen zu verwenden ist (Quadratisch).
Für den Fall, dass historische Daten vorliegen (ohne Versuchsplan) ist evtl. ein quadratisches Modell ohne Wechselwirkungen u.U. kritisch, da die Bestimmung bei korrelierenden Daten problematisch ist.

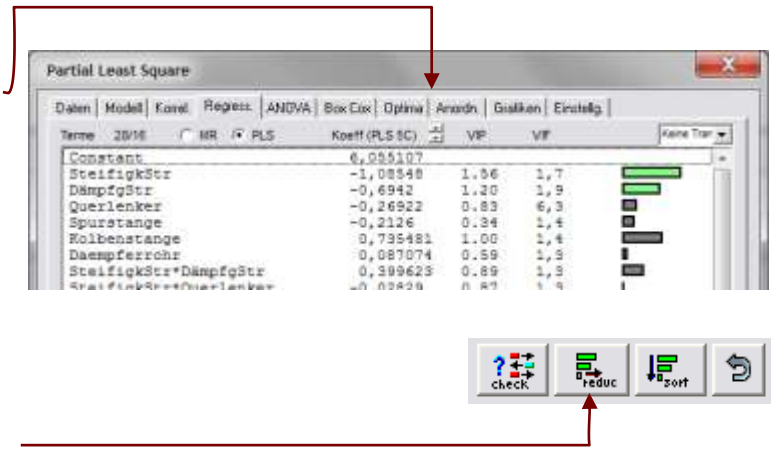


Unter dem nächsten Reiter kann eine mögliche kritische Korrelation erkannt werden. Aus Erfahrungen hat sich gezeigt, dass es ab r > 0,60 Probleme geben kann.



Partial-Least-Square PLS und Kennzahl VIP

Auf der Ergebnisseite, die hier auch Regression genannt wird, sind die Koeffizienten auf Basis von 5 Componenten zu sehen. Gibt es keine Korrelation zwischen den Parametern entsprechen die Zahlenwerte denen der multiplen Regression.



Die VIP Zahl zeigt die „Wichtigkeit“ des Parameters in die Projektion der Componenten. Wie beschrieben, sollte VIP ≥ 0,5 die direkten Parameter und VIP ≥ 1,0 für höherwertige Terme sein. Über die Taste ***reduce*** können diese entsprechende markiert werden und aus dem Modell herausgenommen werden (rotes X).

Aller weiteren Funktionen, insbesondere die Grafiken, entsprechen den Beschreibungen zur multiplen Regression.

Literatur

Taschenbuch der statistischen Qualitäts- und Zuverlässigkeitsmethoden

Die wichtigsten Methoden und Verfahren für die Praxis.

Beinhaltet statistische Methoden für Versuchsplanung & Datenanalyse, sowie Zuverlässigkeit & Weibull.

- Statistische Verteilungen und Tests & Mischverteilungen
- Six Sigma Einführung und Zyklen
- Systemanalysen Wirkdiagramm, FMEA, FTA, Matrizen-Methoden
- Shainin- und Taguchi-Methoden
- Versuchsplanung DoE, D-Optimal
- Korrelations- und Regressionsverfahren
- Multivariate Datenauswertungen
- Prozessfähigkeit – Messmittelfähigkeit MSA 4 und VDA 5
- Regelkarten
- Toleranzrechnung und Monte-Carlo-Simulation
- Statistische Hypothesentests
- Weibull und Lebensdaueranalysen
- Stichprobengröße

190 Seiten, Ringbuch

ISBN: 978-3-00-043678-9